2015

Application of analytic tools for materials selection

Pallavi Dubey
Iowa State University

Application of analytic tools for materials selection

by

Pallavi Dubey

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee
Sigurdur Olafsson, Co-Major Professor
Krishna Rajan, Co-Major Professor
Caroline Krejci

Iowa State University

Ames, Iowa

2015

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Sigurdur Olafsson for his constant support and patience. I am very grateful to him as he has always been there and helped me with my doubts and his useful insights. I am grateful to Professor Krishna Rajan for letting me be a part of his research team and for giving me such a beautiful learning environment and for supporting me throughout. Under his guidance I have got an opportunity to understand the complexities of Material Science and freedom to apply what I have learnt in the field of Industrial Engineering. I would also like to thank my thesis committee member Dr. Caroline Krejci for her time and patience.

I have been fortunate to have worked with such a talented research team. I would like to specifically mention Dr. Scott Broderick for his valuable guidance and for always being there and finding time to help clear my doubts. Without his help and insightful ideas, my thesis would not have been possible.

I would also like to thank my fellow group mates especially Rupa, Kevin and Sri for their support and constant encouragement.

ABSTRACT

The objective of this thesis is the targeted design of new wear resistant materials through the development of analytic frameworks. The building of databases on wear data, whether through calculation or experiment, is a very time-consuming problem with high levels of data uncertainty. For these reasons of small data size and high data uncertainty, the development of a hybrid data analytic framework for accelerating the selection of target materials is needed. In this thesis, the focus is on binary ceramic compounds with the properties of interest as friction coefficient and hardness and with the objective being to minimize friction while improving the wear resistance. These design requirements are generally inversely correlated, further requiring the data science framework that is developed in this thesis.

This thesis develops a new hybrid methodology of linking dimensionality reduction (principal component analysis) and association mining to aid in materials selection. The novelty in this developed approach is the linking of multiple data mining methodologies into a single framework, which addresses issues such as physically-meaningful attribute selection, addressing data uncertainty, and identifying specific candidate materials when property trade-offs exist. The result of this thesis is a hybrid methodology for material selection, which is used here for identifying new promising materials for wear resistant applications.

CHAPTER 1

INTRODUCTION

A challenge in wear applications is the dual requirements of low friction combined with high wear resistance. A particular application for wear resistant materials is as a coating for metals, with the coating typically being a ceramic material. The difficulty however is in the time-consuming collection of wear data, whether through computation or through experiment. This challenge has resulted in a small existing data, which results in design difficulty. A further application of this class of wear resistant materials is for lubricants which are used to achieve low friction; example applications include in high temperature environments, where an improvement in the hardness of the material is required [1].

1.1 Objectives and Novelty of Work

When a large data size exists, identifying the target region and property correlations is straightforward. However, when few data exist, identifying physically significant relationships to guide the selection of next material candidate is difficult. This is especially problematic when the data collection on these candidate materials is time-consuming, as is the case here. Numerous data mining approaches exist for objectives ranging from dimensionality reduction, regression, uncertainty quantification, and defining associations; however, given the small data size solely using the approaches developed is not sufficient. Rather, a hybrid approach, which judiciously utilizes specific aspects of each technique, is required. This thesis develops such a hybrid approach by combining these various

approaches into a new methodology, which starts from small data and poorly defined physics to the identification of design rules for accelerated material selection.

A general correlation between hardness and friction coefficient exists (Fig. 1.1). The property target is high hardness and low friction coefficient. Moving into the targeted region will expand the use of these wear materials to high temperature applications. In our study of correlation amongst physical and engineering properties, we take advantage of the ability of data mining methods to screen the properties of different materials when the related data points are small in comparison to independent variables. The impact of this work includes the development of classification rules and prediction models for developing reduced order models. In other words these informatics-based techniques can be used to serve as a means for estimating parameters when data for such calculations are not available.

Using principal component analysis (PCA), partial least square (PLS) regression, Correlation based feature selection (CFS) subset evaluation method and classification apriori algorithm we have derived a method to examine a dataset which has very less data points in comparison to independent variables. These various approaches are discussed in the next section. By piecing different data mining techniques together we have made an attempt to understand the physics behind what makes a material harder and allows it to have less friction at the same time. This work has similar objectives to other approaches, which try to identify trends between material descriptors and properties [18,19]. However, in those works, identifying the key attributes and identifying trends in the data leads only to the empirical mapping of known data. The novelty contributed by the

approach developed here is that by integrating these aspects with predictive and associative algorithms, we convert these mappings into a selection map encompassing unknown materials as well, thereby defining the target candidates.

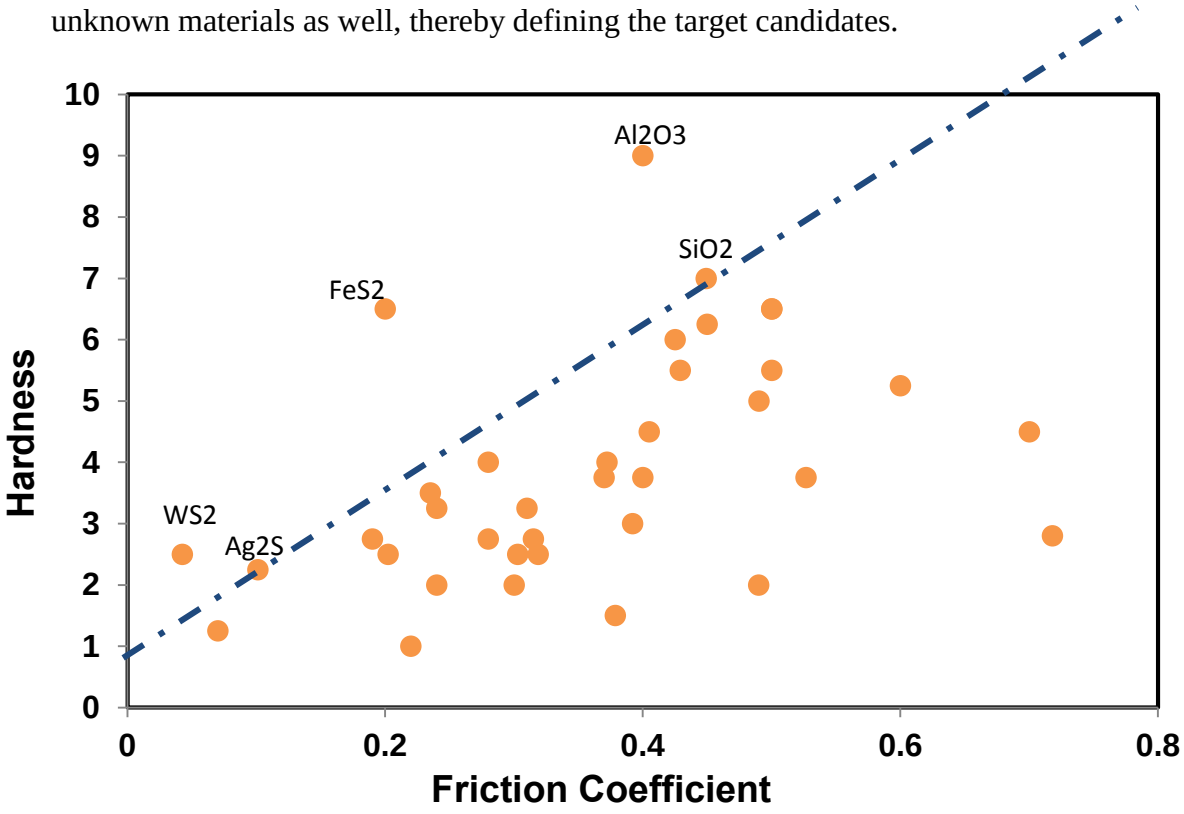

Figure 1.1 The relationship between hardness and friction coefficient. The objective is to increase hardness and decrease friction coefficient, although a boundary in the design of these materials is present in the existing data.

This research aids in understanding how independent variables contribute to the prediction of the engineering properties, particularly when the data is small and sparse with high levels of uncertainty. Different methodologies on attribute selection, and particularly understanding each aspect of these methodologies, are explored and linked to predictive

approaches for developing a quantitative structure-property relationship (QSPR) and developing a "virtual" material library. This virtual library was developed from the small knowledge base. The hybrid informatics approach results in increasing by four times the knowledge base. This thesis also focuses upon comparing different data mining techniques and to address issues such as over fitting, robustness, and uncertainty. Approaches in analytic tools and association mining are then further utilized and integrated with this new approach for selecting the best candidates when an explosion in data size occurs.

1.2 Data Mining

This thesis explored and applied multiple data mining techniques, with aspects of the following primarily utilized: principal component analysis (PCA), partial least squares (PLS), CFS subset evaluation and *a priori* classification using Class Association Rules (CARs). Future work will use qualitative decision analysis methods to identify the compounds with desired balance of the properties of wear resistance defined by the classification rules. The two properties of friction coefficient and hardness are considered in this thesis but an approach to how this can be applied to more than two properties is also addressed.

PCA [2-6] is a projection technique for handling multi variable data that consists of interrelated variables. It inherently decomposes the covariance (or correlation) matrix by calculating the eigenvalues and eigenvectors of the matrix. This decomposition helps in reduction of information dimensionality. As we are selecting only important attributes

through this method, irrelevant and some relevant information is lost but at the same time this method does makes sure to minimize the loss of information and maximize the variance of the linear combination of the variables and uncorrelated axes leading to the transformation (i.e rotation) of the original coordinate system. The constructed axes, referred to as principal components (PCs) correspond with eigenvectors of the original data covariance matrix and are orthogonal to each other. They consist of loadings, which are the weights for each original variable and scores containing information of original samples in a rotated coordinate system. Although the number of PCs equals the number of dimensions of the original data, a few PCs are usually sufficient to capture the major information from the data defining the system. PCA is a powerful tool for understanding the underlying physics within materials science problems and has been used to address materials science issues for a variety of reasons and materials [7-11].

PLS [12-17] is used to make the QSPR model for the given data. PLS has an advantage over typical linear regression techniques of handling co linearity among properties and missing data. As PCA is an analysis for one data matrix. Multivariate regression is for correlating the information in one data matrix to the information in another matrix. PLS is one way to do multivariate regression. Typically one matrix is a cheap measurement of some sort and the other matrix with which we are correlating it can be either very expensive, difficult to measure or time consuming. So this method is used to predict the expensive matrix with the help of the cheap one. Like PCA, in PLS the data is converted to a data matrix with orthogonalized vectors. The relationship discovered in the dataset (training data) can then be applied to a test dataset based on the differences in known properties appearing in both the training and the test sets. The accuracy of prediction model

improves with increasing number of conditions and responses, and thus all predictions shown in this paper can improve with large dataset including more systems and more properties/parameters [2].

CFS subset evaluation is another method of attribute selection. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Also exhaustive search was done for this evaluation, as it performs an exhaustive search through the space of attribute subsets starting from the empty set of attributes and reports the best subset found. Then classification of the reduced dataset was performed with the help of apriori algorithm using class association rule (CARs). A classification data set is in the form of relational table, which is described by a set of distinct attributes (discrete and continuous), whereas association algorithm cannot be performed on a continuous dataset. So we first discretize each continuous attribute. After discretization, we can then transform each data record to a set of (attribute, value) pair of an item. These rules helped in identifying the little nuggets of insight in the data. By calculating the confidence, support and the lift values for each rule we did end up getting six very good rules as are discussed in the 5^{th} chapter, which can help and contribute in the further analysis of the properties of wear resistance and how to improve them. Then there is a future work presented using qualitative decision analysis method to identify the compounds that best satisfy the classification rules.

Similar work taking binary compounds into consideration has been done but the work in the field of wear resistance application and studying the peculiarities of hardness and friction coefficient properties to see what physical properties affect these engineering

properties and how decision analysis based on these properties can help in a better wear resistance application has not been explored earlier. Also a new methodology and approach to materials development using data mining and qualitative decision theory techniques has been introduced in this thesis. It also provides a formal way to handle imprecision and inaccuracies inherent in material properties predicted by machine learning algorithms. This thesis also demonstrates how data mining and decision theory can complement each other in the overall process of materials development and optimization. The methods explored in this thesis will also help us in two ways one, it is applicable to material selection, and two it can be applied as an inverse problem of identifying promising applications for new materials [19]. Also to come up with the combination of techniques to tackle the problem of analyzing the data when the independent variables are much more in comparison to the data points, hence the chances of over fitting a model are very likely. To make choices in this direction we need to look into some relevant observations and deconstruct those observations, and for this we need a model. There are two prediction models and six classification rules as a result of this thesis, which have helped the material scientists, explore the physics behind these two engineering properties further.

1.3 Thesis Outline

This thesis is organized as shown in Figure 1.2, addressing applications of data mining for the development of materials through engineering properties based on their physical properties.

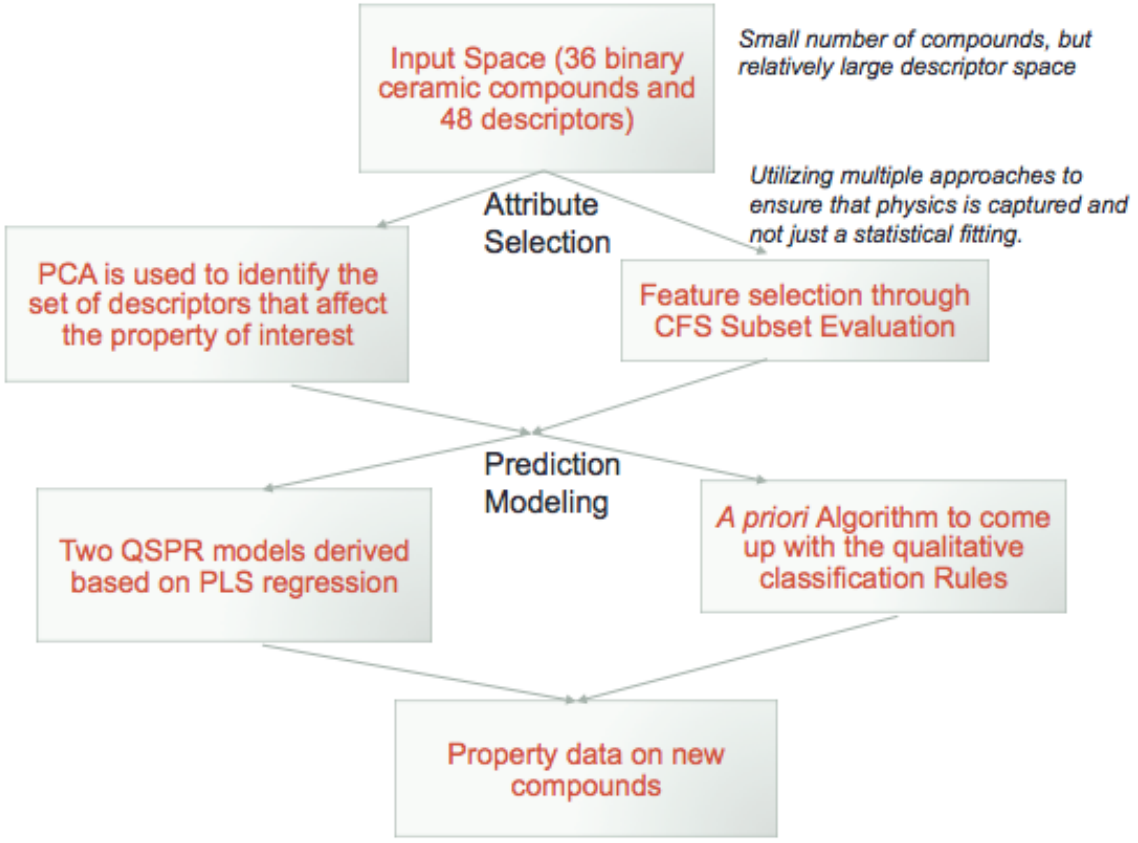

Fig 1.2 The logic of this thesis, with chapter 2 and 3 dealing with the application of PCA and PLS data mining techniques, Chapter 4 with QSPR model application on the virtual data set, Chapter 5 with classification technique and Chapter 6 has a proposal for application of qualitative decision analysis with data mining technique of classification.

In chapter 2, we will be discussing the logic of the PCA technique and its application on the data set of 36 compounds, showing how data mining can be used to reduce the number of parameters and the results are showing which attributes play an important role in describing the hardness and friction coefficient of a material. We will also discuss the constraints and the reasoning behind selecting only a certain important attributes out of the

total result. In chapter 3, I demonstrate how data mining can be used to predict these two important properties of wear resistance and the logic behind PLS and have then discussed the results, leading to a QSPR model.

In chapter 4, Development of the virtual database has been discussed and the application of the QSPR model on the data set has been done to evaluate the results and hence, the model.

In chapter 5, Development of classification rules and another approach of feature selection (i.e CFS subset evaluation) has been discussed. Also the comparison of both the results have been done in this chapter.

Chapter 6 summarizes the work and makes suggestions as to the future direction of this work and the implications it has on the development on new materials as well as new applications with such requirements.

1.4 References

1. P. Menezes, M. Nosonovsky, S. P. Ingole, S. V. Kailas and M. R. Lovell. Tribology for Scientists and Engineers: From Basics to Advanced Concepts, Springer 2013th, New York, 2013.

2. Scott R. Broderick, "Statistical learning for alloy design from electronic structure calculations," (PhD diss, Iowa State University, 2009), 1.

3. Suh C, A. Rajagopalan, X. Li, K. Rajan. The application of principal component analysis to materials science data. DATA Sci. J. 2002;1:19

4. Daffertshofer A, Lamoth CJC, Meijer OG, Beek PJ. PCA in studying coordination and variability: a tutorial. Clin. Biomech. 2004;19:415.

5. Ericksson L, Johansson E, Kettaneh-Wold N, Wold S. Multi- and Megavariate Data Analysis: Principles, Applications. Umea: Umetrics Ab, 2001.

6. Berthiaux H, Mosorov V, Tomczak L, Gatumel C, Demeyre JF. Principal component analysis for characterising homogeneity in powder mixing using image processing techniques. Chem. Eng. Process. 2006;45:397.

7. Nowers JR, Broderick SR, Rajan K, Narasimhan B. Combinatorial Methods and Informatics Provide Insight to Physical Properties and Structure Relationships with IPN Formation. Macromolecular Rapid Communications 2007;28:972.

8. Rajagopalan A, C. Suh, X. Li, K. Rajan. "Secondary" Descriptor Development for Zeolite Framework Design: an informatics approach. Applied Catalysis A: General 2003;254:147.

9. Sieg SC, Suh C, Schmidt T, Stukowski M, Rajan K. Principal component analysis of catalytic functions in the composition of heterogeneous catalyst. QSAR & Combinatorial Science 2007;26:528.

10. Suh C, Rajan K. Combinatorial Design of Semiconductor Chemistry for Bandgap Engineering. Applied Surface Science 2004;223:148.

11. Broderick S, Suh C, Nowers J, Vogel B, Mallapragada S, Narasimhan B, Rajan K. Informatics for Combinatorial Materials Science. JOM Journal of the Minerals, Metals and Materials Society 2008;60:56.

12. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 2001;58:109.

13. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. vol. 18, 2002. p.39.

14. Rosipal R, Kramer N. Overview and recent advances in Partial Least Squares. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J, editors. Subspace, Latent Structure and Feature Selection Techniques. Berlin/Heidelberg: Springer, 2006. p.34.

15 Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Analytica Chimica Acta 1986;185:1.

16. de Jong S. SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 1993;18:251.

17. Phatak A, Jong SD. The geometry of partial least squares. vol. 11, 1997. p.311.

18. Yousef Saad, Da Gao∗, Thanh Ngo, Scotty Bobbitt, James R. Chelikowsky, Wanda Andreoni. Data mining for materials: Computational experiments with AB compounds. November 28, 2011.

19. P Sirisalee, G T Parks, P J Clarkson and M F Ashby. A new approach to multi-criteria material selection in engineering design. International conference on engineering design ICED 03 Stockholm, 2003.

20. H. Liu and H. Motoda. Feature selection for knowledge discovery and data mining, Kluwer Academic, Massachusetts, 2000.

21. John, G., Kohavi, R., and Pfleger, K. (1994). Irrelevant feature and the subset selection problem. In Machine learning: Proceedings of the Eleventh International Conference, pages 121-129. Morgan Kaufmann Publisher.

22. Bing Liu, Yiming Ma, and Ching-Kian Wong. Classification using association rules: weaknesses and enhancement.

23. R. Martinella. Selection and application of wear-resistant materials to increase service life of components. CISE Tecnologie Innovative S.p.A. 20090, 1993.

24. A. Fischer. Well-founded selection of materials for improved wear resistance. Wear 194 (1196) 238-245.

25. Sorokin G. M. and Malyshev V. N. Criterion of wear resistance for ranking steels and alloys on mechanical properties. International Journal of Material and Mechanical Engineering 2012;1:6.

26. I.I. Garbar. Structure-based selection of wear-resistant materials. Wear 181-183 (1995) 50-55.

27. E. W. Bucholz, C. S. Kong, K. R. Marchman, W. G. Sawyer, S. R. Phillpot, S. B. Sinnot, K. Rajan. Data-driven model for estimation of friction coefficient via informatics methods. Tribol Lett (2012) 47:211-221

CHAPTER 2

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis provides a tool for visualizing and quantifying relations between many variables. This is done through bi plots (which have both scores as well as loadings plot) as it gives us the description of both the independent variables and samples. Score plots are used for outlier detection; even though the PCA describes the common phenomenon in the data and not individual peculiarities, through outlier detection and removing those outliers, gives us the major part of the data which can then be used for pattern recognition. Loadings plot provide us the information about the variables, for example it can be used to explore the reasons what make a sample an outlier. The equation X=TP'+E where X is the data matrix, T are the scores, P are the loadings (hence its transpose is used in the equation) and E is the residual i.e. the unexpected part of the data or the noise in the data. Each Principal component consists of one score and one loading vector. Component one which is the first component of the resultant matrix TP' has highest possible variance, and next highest is of the component orthogonal to the 1^{st} component and so on and so forth.

Through PCA calculations we then calculate the k^{th} variable to see which attributes are most relevant and hence this methodology is used here as a dimensionality reduction approach.

2.1 Mathematics of PCA

For a thorough explanation of PCA, the treatment from different sources are combined here [7]. Let us consider the case of a vector x of p number of variables. With

$\alpha_1^T = [\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1p}]$, the variance of the linear function $z_1 = \alpha_1^T x$ is maximized in PCA. The linear function, $z_2 = \alpha_2^T x$ which is uncorrelated with $z_1 = \alpha_1^T x$, can then be calculated to capture the remaining variance. Therefore the *k*-th linear function, $z_k = \alpha_k^T x$, is calculated to have maximum variance and to be uncorrelated with $\alpha_1^T x, \alpha_1^T x, \cdots \alpha_{k-1}^T x$. Consider the case where the vector of random variables *x* has a known covariance matrix *S*. α_k is an eigenvector of covariance matrix *S* corresponding to its *k*-th largest eigenvalue λ_k. If α_k is chosen to have unit length ($\alpha_k^T \alpha_k = 1$), then the variance of z_k is $\text{var}(z_k) = \lambda_k$. To populate the first projection vectors α_1 in $z_1 = \alpha_1^T x$, PCA finds maximum variance, such that

$$\alpha_1 = \underset{\alpha_1^T \alpha_1 = 1}{\arg\max}[\text{var}(\alpha_1^T x)] = \underset{\alpha_1^T \alpha_1 = 1}{\arg\max}[\alpha_1^T S \alpha_1] \tag{2.1}$$

With the constraint of unit length of α_k and maximum variance of z_1, the method of Lagrange multipliers can be applied as

$$\max(L) = [\alpha_1^T S \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1)] \tag{2.2}$$

Where λ is a Lagrange multiplier. Since differentiation gives the maximum value, equation (A.2) results in

$$(S - \lambda I_p)\alpha_1 = 0 \tag{2.3}$$

Where I_p is a ($p \times p$) identity matrix. This is known as the problem of eigenstructure for the covariance matrix. To avoid a trivial null solution, (S- λI_p) should be zero. λ and α_1 should be an eigenvalue of *S* and the corresponding vector respectively. Therefore, the eigenvalue λ represents the variance because:

$$\text{var}(\alpha_1^T x) = \alpha_1^T S \alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda \tag{2.4}$$

Since variance should be maximized in PCA, the eigenvalue λ must be as large as possible. The vector α_1 is the eigenvector corresponding to the largest eigenvalue λ_1 of *S*. A graphical

representation of the eigenvectors and eigenvalues and the assignment of PCs is shown in Figures A.2 and A.3. The second principal component maximizes the variance.

$$\alpha_2 = \arg\max_{\alpha_2^T \alpha_2 = 1} [\alpha_2^T S \alpha_2] \tag{2.5}$$

Subject to the constraint, $\mathrm{cov}(\alpha_1^T x, \alpha_2^T x) = 0$. Thus, it should be uncorrelated with $z_1 = \alpha_1^T x$. Using the method of Lagrange multipliers,

$$\max(L) = [\alpha_2^T S \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \phi(\alpha_2^T \alpha_1 - 0)] \tag{2.6}$$

Where λ and ϕ are Lagrange multipliers. The following relations result in $(S - \lambda I_p)\alpha_2 = 0$. The vector α_k is called the loadings for the *k*-th principal component (PC). The algorithms for calculation of principal components are mainly based on the factorization of matrices. Singular vector decomposition (SVD) and eigenvalue decomposition are the main techniques for factorization of matrices. For any (*I*×I) matrix A and *P* which are non-zero orthonormal matrices, the eigenvalue problem can be expressed as

$$AP = P\Lambda \tag{2.7}$$

Where Λ is an eigenvalue matrix and its components are $\Lambda = diag\{\lambda_1, \cdots, \lambda_I\}$. Then matrix A by eigenvalue decomposition is

$$A = P\Lambda P^T = \sum_{i=1}^{I} \lambda_i p_i p_i^T \tag{2.8}$$

Here, the property $P^{\mathrm{T}}=P^{-1}$ was used from the fact that *P* is orthonormal. If a covariance matrix *S* of *X* is a matrix *A*, the data manipulation involves decomposition of the data matrix *X* into two matrices *V* and *U*, and *V* is orthonormal,

$$S = X^T X = VU^T UV^T = V\Lambda V^T \tag{2.9}$$

The columns of *U* are known as scores and those of *V* are called loadings. PCA is a technique to decompose eigenvalues of a covariance matrix, *S*, of a given data matrix. The loadings can be understood as the weights for each original variable when calculating the principal components. The matrix *U* contains the original data in a rotated coordinate

system. The mathematical analysis involves finding these new "data" matrices *U* and *V*. The dimensions of *U* (i.e. its rank) that capture all the information of the entire data set of *X* (i.e. # of variables) is far less than that of *X* (ideally 2 or 3). One now compresses the N dimensional plot of the data matrix *X* into 2 or 3 dimensional plot of *U* and *V*. While the eigenvalues geometrically represent the length of each of the principal axes (i.e. scores), the eigenvectors of the covariance matrix represent the orientation of principal axes of the ellipsoid (i.e. loadings). By using just a few latent variables, the dimensionality of the original multivariate data sets are reduced and visualized by their projections in 2D or 3D with a minimal loss of information. Therefore, PCA is a process of dimensionally reduced mapping of a multivariate data set [2-6].

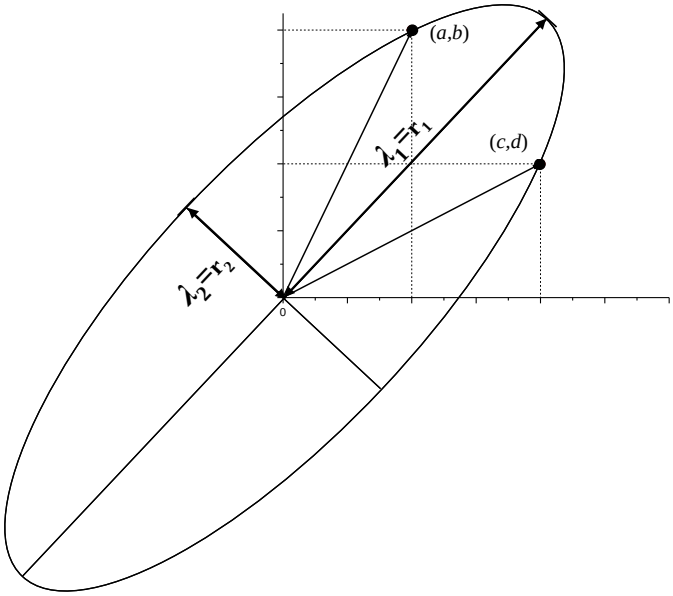

Figure 2.1. A graphical representation of the data points and their eigenvalues

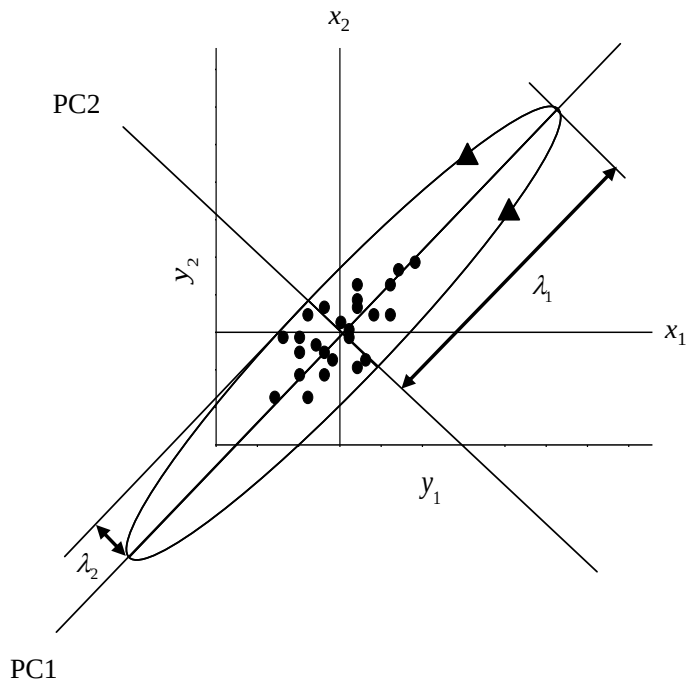

Figure 2.2. Determination of two principal components (PC1 and PC2) in a new scaled coordinate, x_1 and x_2

2.2 Results of the PCA Analysis

As described above the first PC accounts for the maximum variance (eigenvalue) in the original dataset, while the second PC is orthogonal (i.e., uncorrelated) to the first and accounts for most of the remaining variance. So after applying the PCA on the multivariate data for dimensional reduction, we get the major pattern of the data while maximizing the variability contained within the dataset [8-11]. From the PCA results (refer appendix) we find that the first 4 PCs capture 81.8% of the total variance within the original data matrix. Individually PC1 captures 39.71%, PC2 captures 19.81%, PC3 captures 15.36% and PC4

captures 6.92%. It also implies that the first two PC axes already reflect almost 60% of the information of the original data of 36 variables for the data on hardness and friction coefficient. The following score plots in Fig. 2.1 (Hardness) and Fig. 2.2 (Friction) shows the interrelationships between the samples within the dataset relative to the first and second PCs.

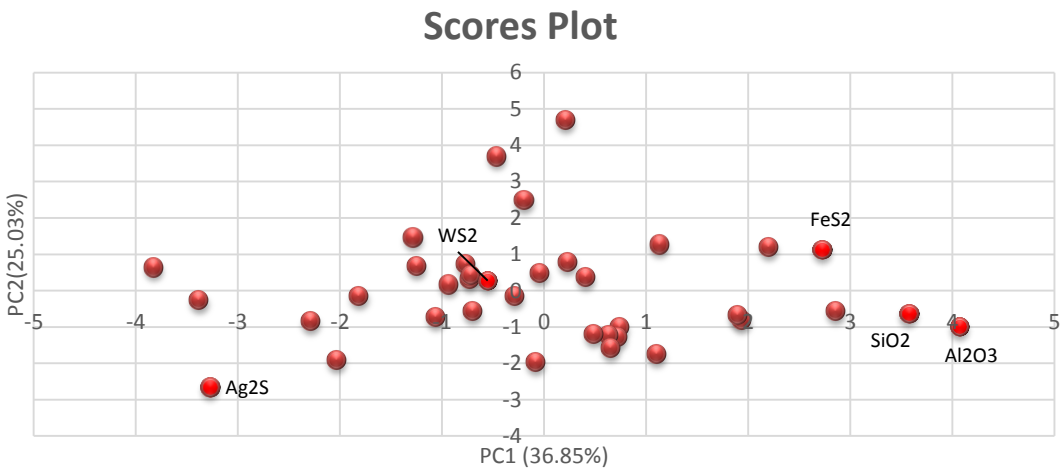

Fig 2.3 Principal component analysis scores plot for the hardness data covering 61.88% of the information of the original data. The best current materials in terms of hardness / friction combination are labeled on this figure.

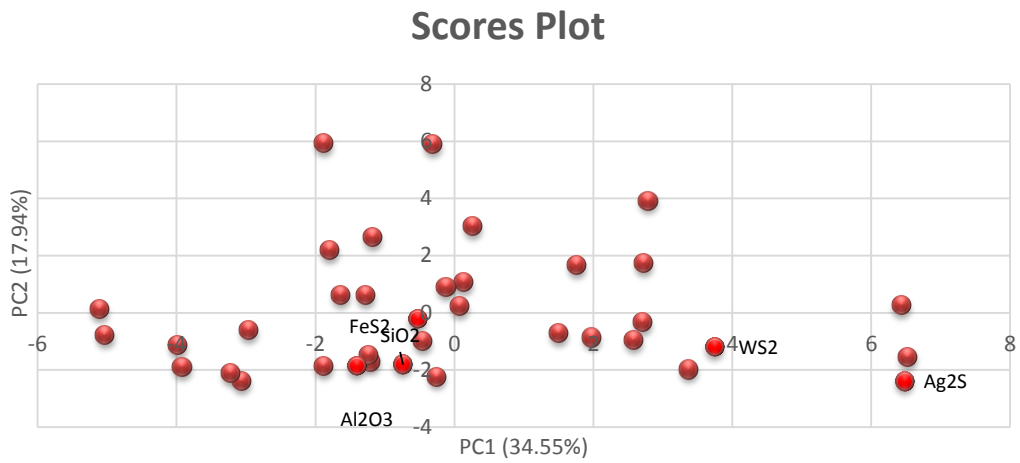

Fig 2.4 Principal component analysis scores plot for the hardness data covering 52.49% of the information of the original data. The best current materials in terms of hardness / friction combination are labeled on this figure.

PCA was used to assess the correlation between each of the descriptors input into the regression analysis and the properties of interest (hardness and friction coefficient). The loadings plot reveals the significance of the different input variables with respect to the target variable, which are our two engineering properties. Also this method helps in checking the outliers using the scores plot and the loadings plot which helps further in the analysis of those outliers. There were no outliers in our analysis so for the next part we have considered the same data.

2.3 Analysis of Variable Importance

To select the descriptors that best define the average hardness and average friction coefficient respectively, we used variable importance in the projection method. In partial least square regression, the relative contribution of each parameter is evaluated using the measure of VIP [8]. Suppose t stands for the target (a particular site of a specific compound), k for the descriptor, r represents the number of descriptors and by P_x we mean xth PC.

$$\text{Importance of the kth variable} = (P_1^K P_1^t + P_2^K P_2^t / \sum P_1^K P_1^t + P_2^K P_2^t) * 100 \quad (2.10)$$

Where P_1^K is the kth component of the first eigenvector (PC1) corresponding to the kth variable and P_1^t is component of PC1 corresponding to the target vector t.

The cutoff of variable importance parameter value is greater than 10 for hardness and 25 for Friction coefficient; these values will be selected as a model parameter. After performing PCA and assessing the correlation of the descriptors and performing regression analysis (which is explained in the next chapter), the results of the analysis can then be compared with the predictive models to understand the physics and limitations of the models.

2.4 Results of Variable Importance

The objective of the attribute analysis is comparing the different models. The results show that there were around eight variables that are above the defined cut off value of 10 for hardness but only four attributes were considered for the QSPR model. The other few attributes that pass the cut off were not considered as they did not improve the accuracy

when added to these selected attributes as shown in the figure, so we concluded that these are the best possible descriptors as they gave the best accuracy.

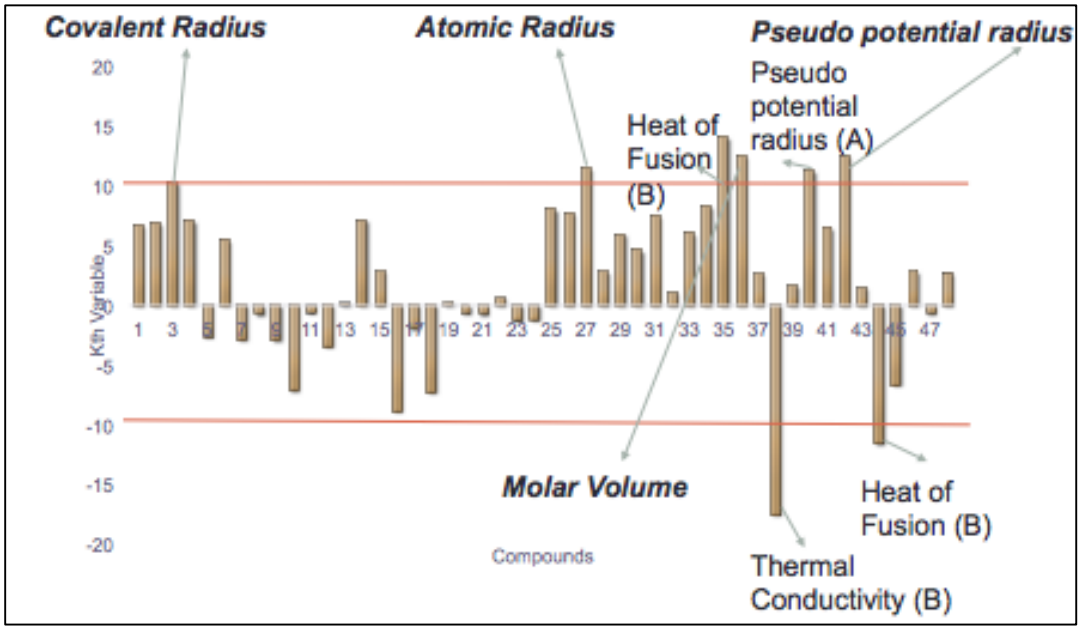

Fig 2.5 Ranking of importance of features on hardness through PCA calculated according to the equation of importance of kth variable. The four most important predictor variables that are giving highest accuracy are Molar Volume, Covalent Radius, Atomic Radius and Pseudo potential radius. The following table shows the variables that correspond to each number in the above graph.

Table 1: Describing each number relating to which property in both graphs showing variable of importance results.

1	Covalent radius (A)
2	Covalent radius (B)
3	Covalent Radius (A+B)
4	Melting point (A) (K)
5	Melting point(B)
6	Melting point (A+B)
7	First Ionization Potential (A)
8	First ionization potential(B)
9	First Ionization Potential(A+B)
10	Martynov-Batsanov electronegativity X [(eV)1/2] A
11	Martynov-Batsanov electronegativity X [(eV)1/2] B
12	Martynov-Batsanov electronegativity X(A+B)
13	Valence electron number, Nv (A)
14	Valence electron (B)
15	Valence electron number (A+B)

Table 1. (Continued)

16	Specific heat A
17	Specific heat B
18	Specific heat (A+B)
19	Pauling electronegativity A
20	Pauling electronegativity B
21	Pauling electron negativity (A+B)
22	Heat capacity A
23	Heat capacity B
24	Heat capacity A+ B
25	Atomic radius A
26	Atomic radius B
27	Atomic radius (A+B)
28	Boiling point A
29	Boiling point B
30	Boiling point(A+B)
31	Density A @293K

Table 1. (Continued)

32	Density B @293K
33	Density (A+B)
34	Molar Volume A
35	Molar Volume B
36	Molar Volume (A+B)
37	Thermal conductivity A
38	Thermal conductivity B
39	Thermal conductivity (A+B)
40	Pseudo potential core radii sum.A
41	Pseudo potential core radii sum.B
42	Pseudo potential core radii sum (A+B)
43	Heat of fusion A
44	Heat of fusion B
45	Heat of fusion (A+B)
46	Heat of vaporization A
47	Heat of vaporization B

Table 1. (Continued)

48	Heat of vaporization (A+B)

For the friction data, the analysis indicates eight important variables; again the best combination is of valence electron, first ionization potential, boiling point and heat of vaporization as the accuracy is best with this combination and other variables are not affecting the accuracy in any way when included. The following graph indicates the results as well.

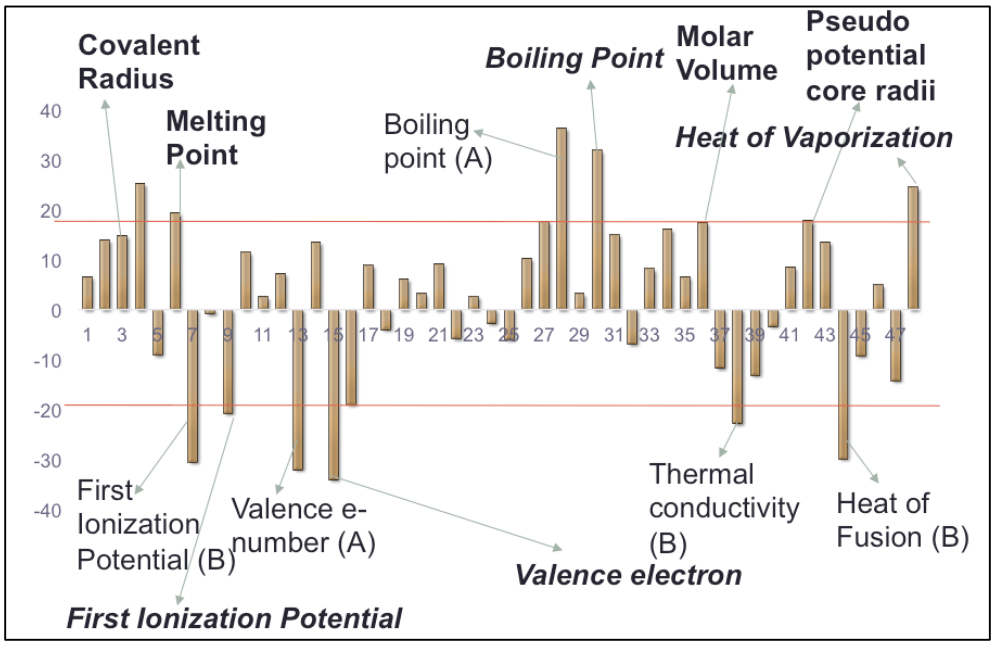

Fig 2.6 Ranking of importance of features on friction through PCA calculated according to the equation of importance of kth variable. The four most important predictor variables that are giving highest accuracy are Molar Volume, Covalent Radius, Atomic Radius and Pseudo potential radius.

2.5 References

1. P. Menezes, M. Nosonovsky, S. P. Ingole, S. V. Kailas and M. R. Lovell. Tribology for Scientists and Engineers: From Basics to Advanced Concepts, Springer 2013th, New York, 2013.

2. Scott R. Broderick, "Statistical learning for alloy design from electronic structure calculations," (PhD diss, Iowa State University, 2009), 99.

3. Suh C, A. Rajagopalan, X. Li, K. Rajan. The application of principal component analysis to materials science data. DATA Sci. J. 2002;1:19

4. Daffertshofer A, Lamoth CJC, Meijer OG, Beek PJ. PCA in studying coordination and variability: a tutorial. Clin. Biomech. 2004;19:415.

5. Ericksson L, Johansson E, Kettaneh-Wold N, Wold S. Multi- and Megavariate Data Analysis: Principles, Applications. Umea: Umetrics Ab, 2001.

6. Berthiaux H, Mosorov V, Tomczak L, Gatumel C, Demeyre JF. Principal component analysis for characterising homogeneity in powder mixing using image processing techniques. Chem. Eng. Process. 2006;45:397.

7. E. W. Bucholz, C. S. Kong, K. R. Marchman, W. G. Sawyer, S. R. Phillpot, S. B. Sinnot, K. Rajan. Data-driven model for estimation of friction coefficient via informatics methods. Tribol Lett (2012) 47:211-221

8. Chong, I.G., Jun, C.H.: Performance of some variable selection methods when

multicollinearity is present. Chemom. Intell. Lab. Syst. 2005;78:103-112

9. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. Chemometrics: A Textbook. Amsterdam: Elsevier, 1988.

10. Jolliffe IT. Principal Component Analysis. New York: Springer-Verlag, 2002.

11. Davis JC. Statistics and Data Analysis in Geology. New York: John Wiley & Sons, 1986.

12. Suh C. Informatics Aided Design of Crystal Chemistry. Engineering Science, vol. Ph.D.: Rensselaer Polytechnic Institute, 2005.

CHAPTER 3

PARTIAL LEAST SQUARE REGRESSION

PLS is one way to do multivariate regression. Principle of PLS is to find components in such a way that their score values have maximum covariance. PCA is for analysis of one data matrix (X). Multivariate regression is for correlating the information in one data matrix (X) to the information in another matrix (Y). Typically the X matrix is a cheap measurement of some sort and the Y matrix may be very expensive/difficult to measure/or Time consuming, so through X we can predict the values of Y by this method.

3.1 Introduction

So from the important descriptors that we get from the PCA are then used for making the prediction model through PLS. To discuss the theory of PLS regression here are two multivariate matrices shown in Fig. 3.1.

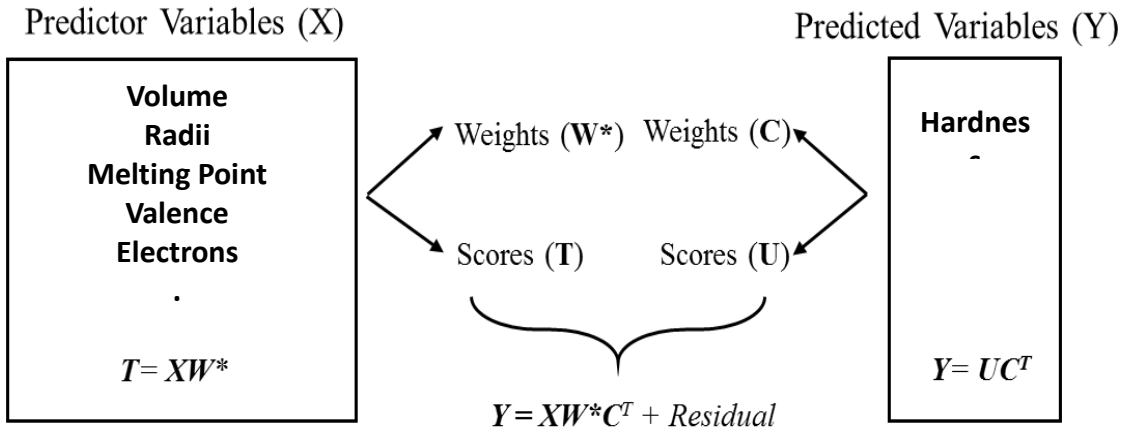

Fig 3.1 Describing through a block diagram the principal of PLS regression method, where X is a cheap matrix Y is the matrix of the data which is difficult to measure and the first

score in X i.e. t1 has the maximum covariance with the first score in Y i.e. u1 and r1 is a constant and where each is a component of a Matrix denoted by capitol letters.

We want to develop a model so in future we don't need both X and Y but just X and then by using X and the model we are able to predict the values of Y. Let us assume that we perform PCA on our Y matrix first. As we know that to perform PCA on a single data matrix is very useful because they have lower rank i.e. they can be described by fewer components than the original number of variables. So if we do a PCA on Y we would then be able to describe our Y matrix in terms its scores time loadings. Y=UQ'+F So through this we conclude that we just have to predict U (i.e. the scores of Y) and through these scores and the loadings we can then predict Y.

In PLS we develop our model in such a way that the first score in X i.e. t1 has the maximum covariance with the first score in Y i.e. u1. So because of this high covariance we can predict the first score in Y by the first score in X so as soon as we have the score values of X we can predict Y. So this the main concept of PLS, it finds components in such a way that their score values have maximum covariance, u1 has maximum covariance with t1 so on and so forth.

PLS does not consist of just doing PCA on X and PCA on Y. Instead of finding the major variation in X and the major variation in Y, PLS looks for a direction in both which is good for correlating X score with the Y score. So it looks for the relevant information (for Y). The mathematics of PLS has been explained in the next section.

3.2 Mathematics of PLS

Partial least squares (PLS) finds the maximum variance in the predictor variables (*X*) and finds the correlation factors between *X* and the predicted variables (*Y)* that have maximum variance. In PLS, two linear combinations are generated from the *X* and *Y* respectively and the maximum covariance between *X* and *Y* is calculated. Consider an *X* matrix of size $N \times K$ and an $N \times M$ matrix *Y*

The following descriptions are mainly based on [1,4,5,7]. The scores of *X*, t_a (a=1, 2, …, A=the number of PLS components) are calculated as linear combinations of the original variables with the weights w^*_{ka}. The mathematical expression is

$$t_{ia} = \sum_k w^*_{ka} x_{ik} \text{ or } T = XW^* \tag{3.1}$$

where k=(1, …, K=the number of *X* variables). The predictor variables, *X*, are expressed as:

$$x_{ik} = t_{i1} p^T_{1k} + t_{i2} p^T_{2k} + \cdots + t_{iA} p^T_{Ak} + e_{ik} = \sum_a t_{ia} p_{ak} + e_{ik} \text{ or } X = TP^T + E \tag{3.2}$$

where e_{ik} is the *X* residuals.

Similarly, for predicted variables *Y*, if the scores of *Y* are u_a and the weights c_{am}:

$$y_{im} = \sum_a u_{ia} c_{am} + g_{im} \text{ or } Y = UC^T + G \tag{3.3}$$

Since scores *X* are good predictors of *Y* in PLS, then:

$$y_{im} = \sum_a t_{ia} c_{am} + f_{im} \text{ or } Y = TC^T + F \tag{3.4}$$

where F represents the error between observed values and the predicted response. Using equation (3.1), the equation (3.2) is also expressed as

$$y_{im} = \sum_a c_{am} \sum_k w^*_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im} \quad \text{or} \tag{3.5}$$

$$Y = XW^{*}C^{T} + F = XB + F$$

From equation (3.3), the PLS regression coefficients β_{mk} is written as

$$\beta_{mk} = \sum_{a} c_{am} w^{*}_{ka} \quad \text{or} \quad B = W^{*}C^{T} \tag{3.6}$$

Geometrically, all the above parameters are shown in Figure 3.1 As discussed before, the multidimensional space of *X* is reduced to the A-dimensional hyper plane. Since the scores are good predictors of *Y*, the correlation of *Y* is formed on this hyper plane. As in PCA, the loadings of *X* (*P*) represent the orientation of each of the components of the hyper plane.

According to the approach of Phatak and de Jong, after *n* dimensions have been extracted the following equations are available.

$$T_n = XW^{*}_n,\ P_n = X^{T}T_n(T_n^{T}T_n)^{-1},\ W^{*} = W_n(P_n^{T}W_n)^{-1} \tag{3.7}$$

The prediction of *y* then has a general form given by equation (2.7)

$$y^{n}_{PLS} = T_n(T_n^{T}T_n)^{-1}T_n^{T}y \tag{3.8}$$

From the equations (3.5) and (2.6), equation (3.6) is written as:

$$\hat{y}^{n}_{PLS} = X\hat{\beta}^{n}_{PLS} = XW^{*}_n(W_n^{T}X^{T}XW^{*}_n)^{-1}W_n^{T}X^{T}X\hat{\beta}_{OLS} \tag{3.9}$$

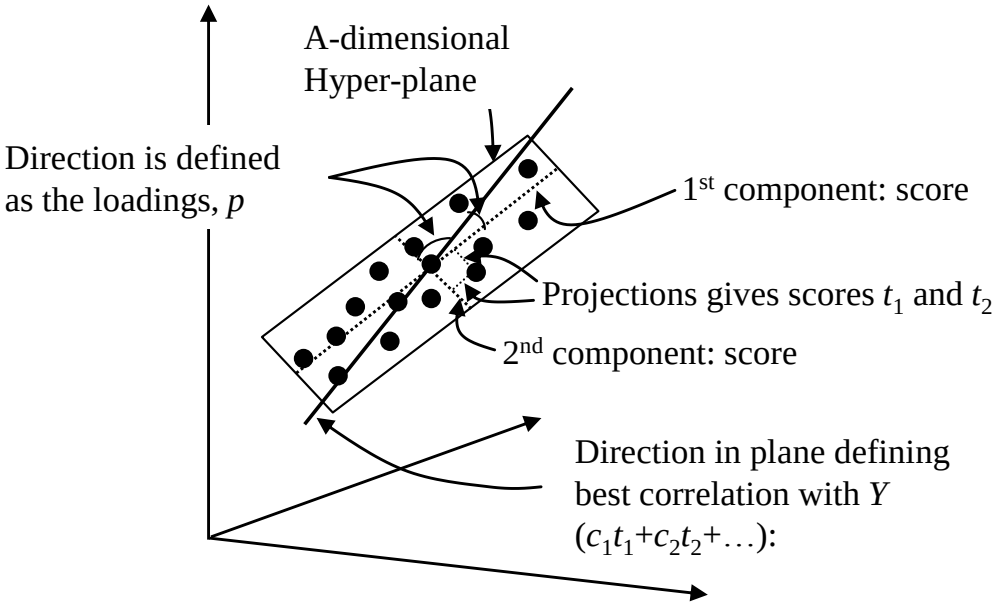

Figure 3.2 Geometrical representation of PLS method [1].

3.3 Results

Before conducting a PLS regression, a multiple R regression was also performed on the data for comparison purposes. The R^2 value for the prediction model of the hardness data was 0.9763; R^2 is the goodness of fit for a formula (usually a straight line) to the data and it support multiple R assumption which in this case supports the selected variables of importance. Multiple R is another measure, which is when greater than 0 and close to 1 state that the two variables are closely related, its value for the hardness data is .98809. Also to state the accuracy of the prediction model the adjusted R-value is calculated, which is .97 for the hardness prediction model. Hence the accuracy of the model is very good. Similarly the R^2 value for the friction prediction model is .8117, Multiple R is .900997 and Adjusted R is .774154. In these cases there can be two theories either the accuracy is too

good due to over fitting or else the model is very accurate. To look into this argument further let's discuss the PLS regression results also.

The Quantitative Structure–Property Relationship Model is achieved after Partial Least Regression, which can then be used for predicting the friction coefficient and hardness of the binary compounds. The following equations were the QSPRs initially developed with the objective of maximizing accuracy and with the results shown in Figs. 3.3 and 3.4.

$$\textit{Hardness} = -1.85*\textit{covalent radius} + 0.928*\textit{atomic radius} - 0.019*\textit{molar volume} - 0.844*\textit{pseudopotential radius} + 2.15 \quad (3.10)$$

$$\textit{Friction Coefficient} = 0.84*\textit{covalent radius} - .00017*\textit{melting point} + 0.021*\textit{atomic radius} + 0.03*\textit{pseudopotential radius} - 0.0084*\textit{heat of vaporization} - 0.347 \quad (3.11)$$

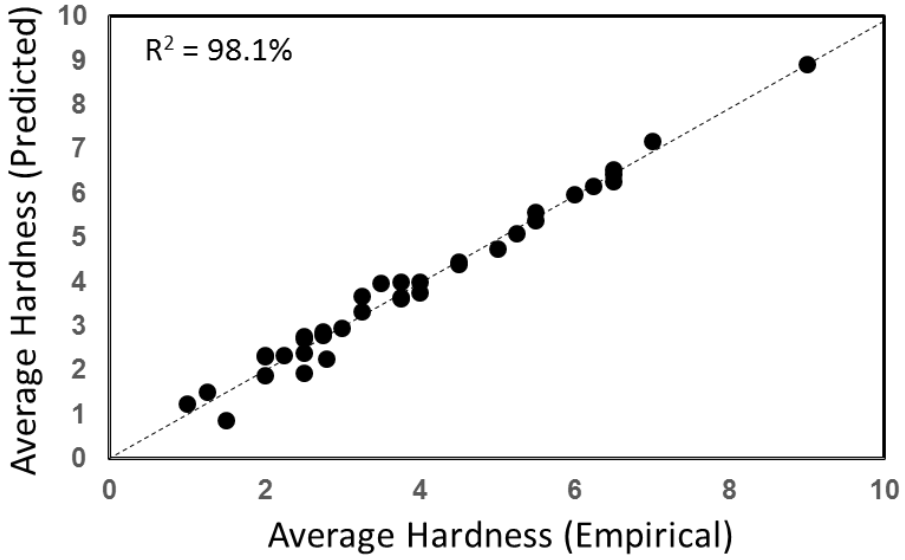

Fig 3.3 Predicted vs. experimental hardness values for the training data through partial least square regression.

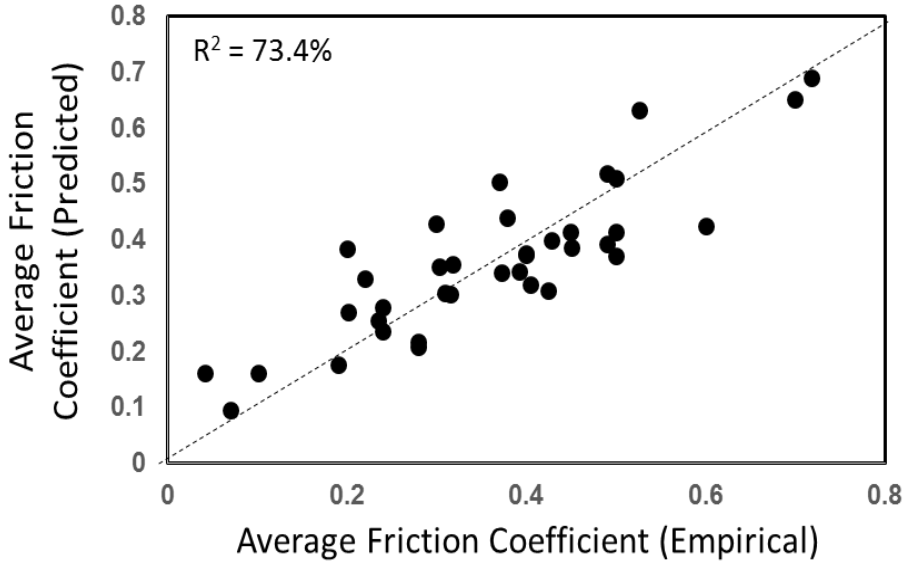

Fig 3.4 Predicted vs. experimental friction coefficient values for the training data through partial least square regression technique.

In developing these models we applied cross validation through the form of leave-one-out (LOO), where a sample is removed when building the model and then is used to test the model accuracy. This is repeated for each sample. This allows us to get the root mean square error (RMSE) and root mean square error of cross validation (RMSECV). The selection of dimensions to include in building the model is such that n is one less than RMSE(n)/RMSECV(n) is equal to unity. Therefore, this demonstrates the typical approach of utilizing PLS where the objective is to maximize accuracy while also employing a cross validation strategy.

However, to further assess the physics of the models, they were applied to "virtual" compounds (these compounds and the associated descriptors are discussed in the next section). The result of applying these QSPRs is shown if Fig. 3.5. Clearly these models, while highly accurate for the training data are insufficient for capturing the physics. The two most obvious issues are the negative friction coefficient values, which is physically unreasonable, and the three outlier chemistries which have hardness over eight times greater than any of the other compounds. This introduces two issues: first, the model is over-fitting the training data to a great extent, where the training data is modeled with high accuracy but has no application to other systems, and second that it is likely over-fit to the physics of the outlier compounds, thus significantly skewing the results. This represents the challenge addressed in this thesis associated with small data sets.

.

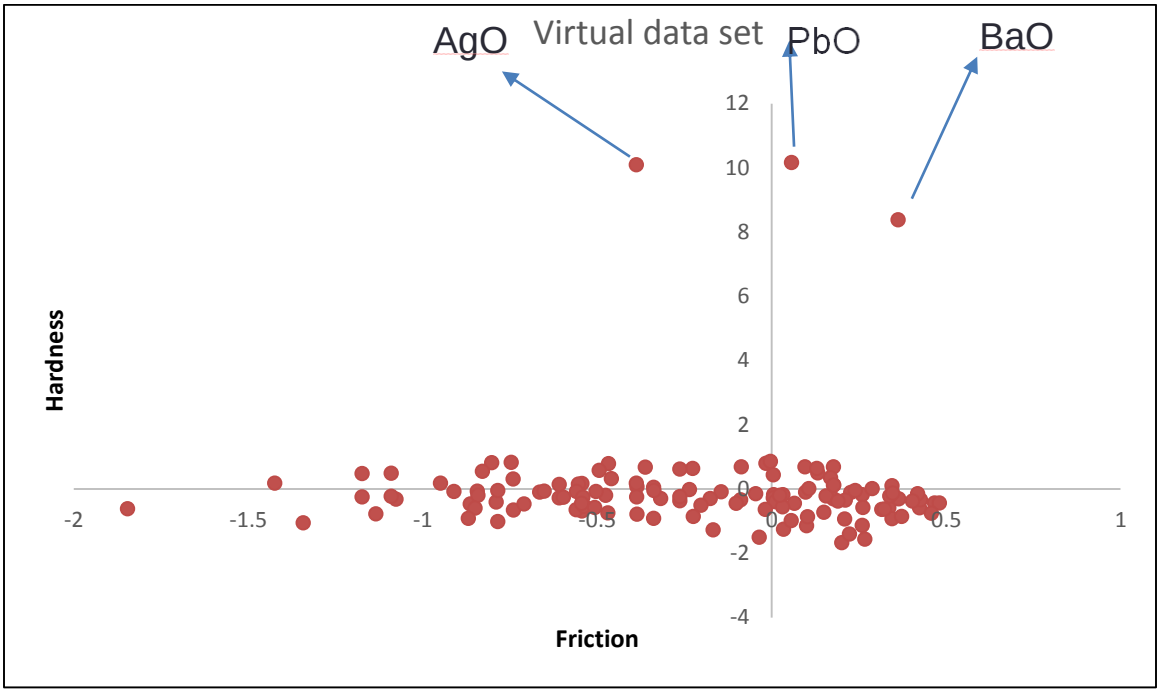

Fig 3.5 The graph between the predicted hardness and predicted friction coefficient after application of the QSPR model on the virtual data set of 135 compounds. The physical unreasonableness of this result demonstrates the challenge with small training data, even with high accuracy and cross-validation employed as is typically done.

To develop a QSPR which is applicable for new systems, the robustness was increased with a trade-off in lower accuracy. This was done by reducing the number of latent variables (LVs), akin to the PCs in PCA, used in the model. Therefore, this introduces less uncertainty and contributions from outliers, as only the LVs describing the general physics are included. This leads to models which fit training data less well, but with much improved fitting to the test data. The result of this new model is as follows:

$$Hardness = .93*covalent\ radius + 0.59*atomic\ radius + 0.14*molar\ volume - 1.85*pseudo\ potential\ radius + 0.44 \quad (3.12)$$

$$Friction\ Coefficient = -.61*Covalent\ Radius - 0.00015*Melting\ Point + 0.019*\ Pseudo\ potential\ core\ radii - 0.0046*Molar\ Volume - 0.19 \quad (3.13)$$

The result of this model, as compared with the input data is shown in Fig. 3.6. In the figure the red squares are the predicted values and the black circles are the measured values for the new prediction model described in the equations 3.12 and 3.13.

The two problems that arose when following the standard approach for developing QSPR on small data is addressed in this updated prediction. The results are all physically reasonable, with the measures falling within the boundaries of the actual data. Further, the model is not over-fitting to outliers, as the predictions, even for outliers in the original data, are clustered with the majority of the other compounds. Therefore, this model is capturing the general guiding physics, without building in physics that is only specific in a small number of cases. While this contributes to potentially missing promising candidate materials which do follow unique physics, it helps ensure that any compounds identified

as having unique properties do indeed have those properties and therefore significantly reduce the number of materials that need to be experimentally explored further.

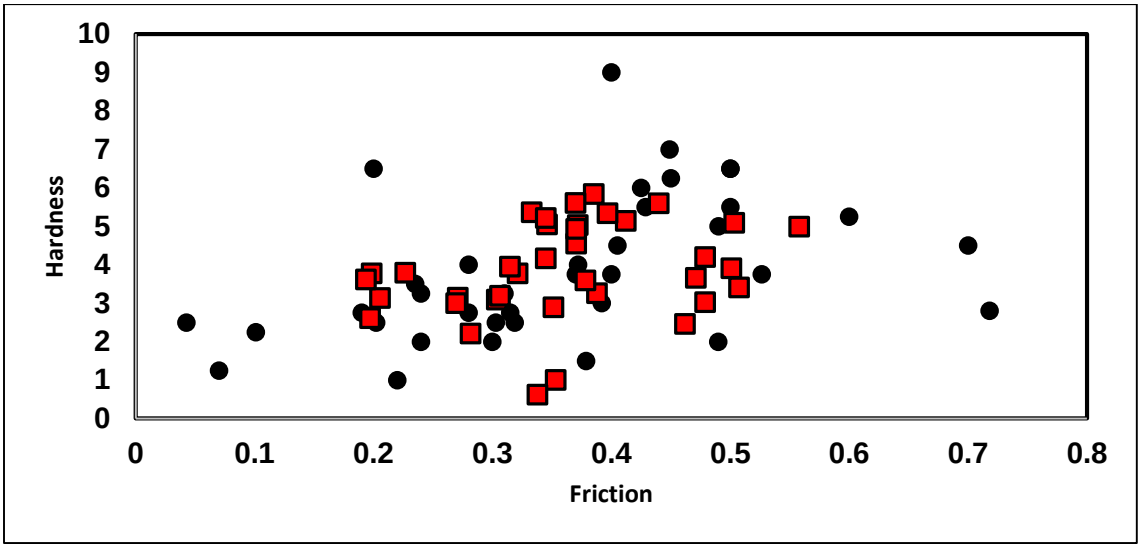

Fig 3.6 The red squares are the predicted values and the black circles are the actual data points. This demonstrates that the new model is not overly impacted by outliers. This ensures higher robustness and the confidence when new compounds with promising properties are identified.

This chapter explored the trade-offs in developing predictive models on small databases. In this chapter, I demonstrated that enhancing the model for robustness at the expense of accuracy leads to more meaningful results, which is further highlighted in the next section. The QSPR modeling is a fast method and can be applied to the system for which you have limited knowledge on. QSPR models can screen the material space much faster than otherwise possible. In the next chapter we will discuss the application of this predictive model on a virtual database and will assess the validity of this model.

3.4 References

1. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 2001;58:109.

2. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. vol. 18, 2002. p.39.

3. Rosipal R, Kramer N. Overview and recent advances in Partial Least Squares. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J, editors. Subspace, Latent Structure and Feature Selection Techniques. Berlin/Heidelberg: Springer, 2006. p.34.

4. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Analytica Chimica Acta 1986;185:1.

5. de Jong S. SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 1993;18:251.

6. Phatak A, Jong SD. The geometry of partial least squares. vol. 11, 1997. p.311.

7. Suh C. Informatics Aided Design of Crystal Chemistry. Engineering Science, vol. Ph.D.: Rensselaer Polytechnic Institute, 2005.

8. E. W. Bucholz, C. S. Kong, K. R. Marchman, W. G. Sawyer, S. R. Phillpot, S. B. Sinnot, K. Rajan. Data-driven model for estimation of friction coefficient via informatics methods. Tribol Lett (2012) 47:211-221

CHAPTER 4

VIRTUAL DATABASE DEVELOPMENT AND ANALYSIS

In this chapter we will predict the hardness and friction coefficient of over 100 new binary compounds for chemistries not existing in our training data and characterized for wear applications. A partial least square data mining approach has been used for the initial analysis of the data. The values of important attributes have been calculated mathematically by normalization using elemental database for the virtual compounds.

4.1 Development of Virtual Database

The experimental data comprises of 36 compounds and their calculated friction and hardness values and mathematically calculated attributes by normalization. All these 36 compounds were binary in nature. Using those 36 compounds we have developed 135 new compounds, combining different elements in a binary form which have chemically possible chemistries. As shown in the following diagram the red squared elements are the cations and the green squared elements are the anions. These elements are the ones of which the 36 compounds were made of. Using these elements and previously untested chemistries but chemically possible combinations we have developed our virtual database.

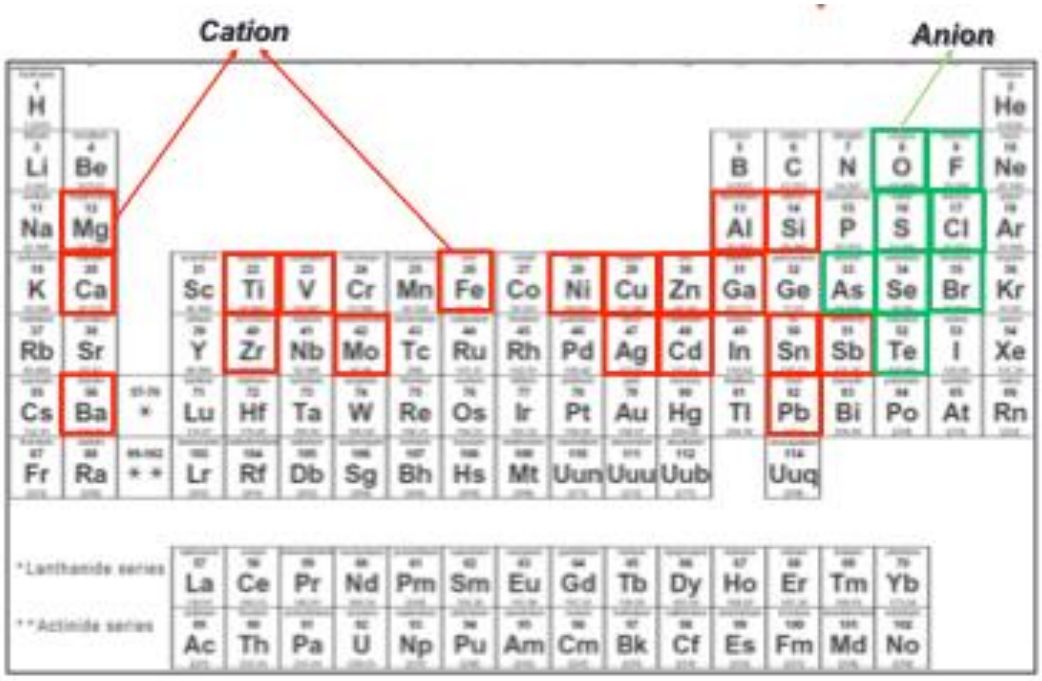

Fig 4.1 Periodic table showing the elements used in the virtual database formation, the red boxes show the elements that act as cations in a binary compound and green are the anions.

The list of compounds and the data calculated have been mentioned in the appendix. For these 135 compounds the data was calculated and then the QSPR model was applied to this database. As discussed in the 1st chapter that we have conducted this analysis to study further those compounds which may lay in the targeted region i.e. the highlighted region of Fig 1.1 for the applications where materials with improved wear resistance performance are required. We also discussed in the 3rd chapter that the accuracy of a prediction model itself cannot validate the model as sometimes it can be an over fitted model. This problem occurs with those models, which have lesser data points in comparison to independent variables. We also conducted a standard method of decision tree analysis, which was again

highly over fitted. To solve this problem further and as described in the previous chapter by improving the robustness of the model we derived new QSPR models. Using these improved models and applying it on the virtual database, the following figure was the derived result.

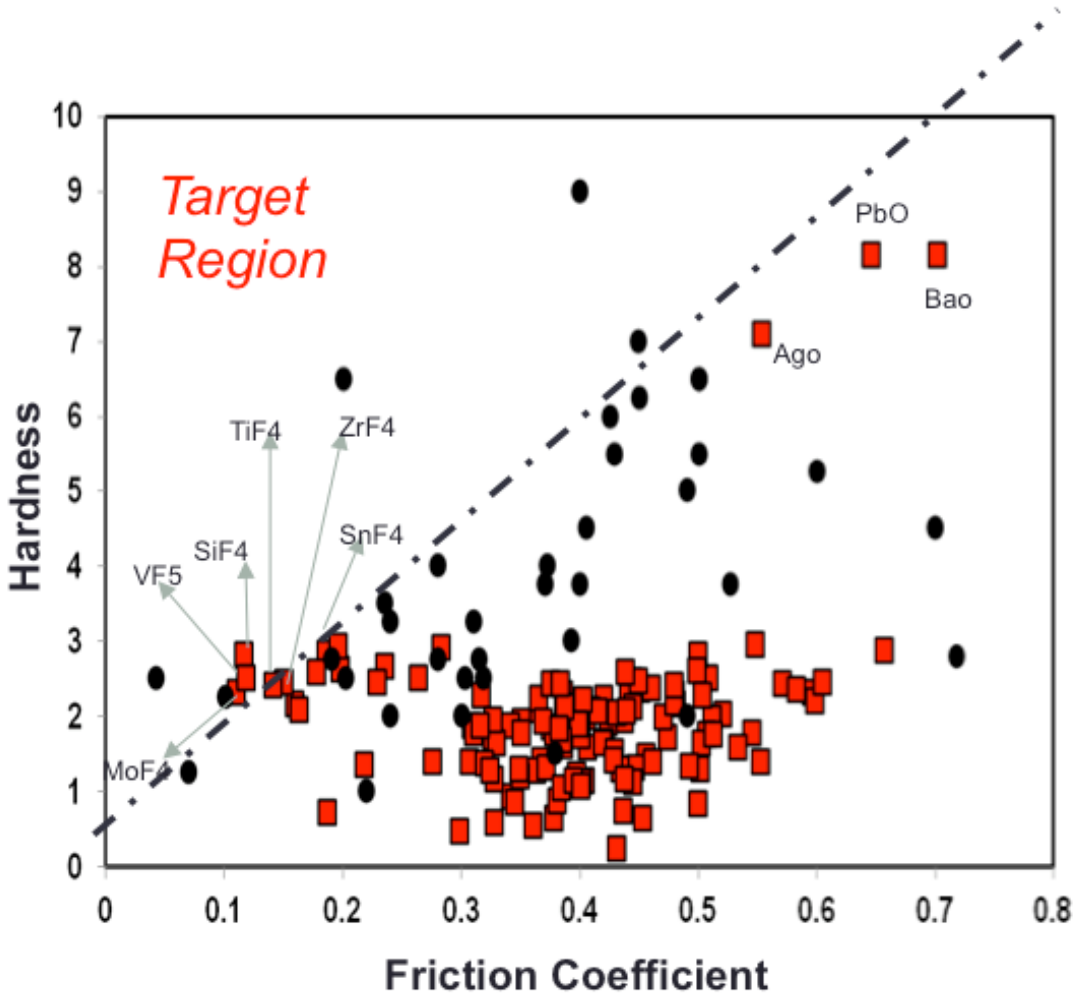

Fig 4.2 The graph between the actual and predicted hardness and friction coefficient after application of the QSPR model on the virtual data set of 135 compounds based on equations 3.12 and 3.13. The red squares are calculated from QSPR model and the black circles are the actual dataset.

As we can see in the graph above the actual dataset (the black circles) has five outliers which are in the highlighted or targeted region, these compounds are SiO_2, Al_2O_3, Ag_2S, WS_2 and FeS_2. Our aim was to explore the virtual database and see if we can find new compounds in the targeted region satisfying the criteria of low friction and higher hardness. We are able to find six compounds with the desired friction and hardness combination. Also there are three new compounds that are highlighted in the graph on the upper corner on the right, which have very high hardness. The following table enlists the compounds with the desired output.

Desired friction and hardness combination	Very high hardness
SiF4, ZrF4, TiF4, VF5, MoF4, SnF4	BaO, PbO, AgO

Table 2: Compounds with desired properties not included in the existing knowledge base.

In the previous chapter, BaO, PbO and AgO were the outlier compounds which had unique physics not present for others. Of note, these compounds are identified from our new model as well. Therefore, although we increased the robust catching more general trends and less sensitive to outliers, are model still captures these three compounds with unique physics. This demonstrates the benefit of this approach for not being overly impacted by outlier compounds while still capturing unique physics when present.

To explore what can be done for studying such type of data further and to get an unbiased predictive model, to help us get useful insights into the physics related to material science another methodology has been explored in the next chapter. There have been different approaches of dealing with dataset of small data points plus more independent variables. I applied classification techniques after applying exhaustive search on the data to come up with association rules which do not manipulate the data but analyze it as it is and I came up with classification rules for hardness and friction coefficient where how these can be made better and how we can compare the virtual dataset on the basis of these rules was found out. Also how to conclude which virtual compound is better than another by defining specific confidence and support has been explored and discussed thoroughly in the next chapter.

CHAPTER 5

DEVELOPMENT OF CLASIFICATION RULES

Apart from PCA we have also looked into another feature selection method in this chapter, which is CFS subset evaluation. In this chapter we will also discuss the two approaches of data mining heuristic and exhaustive and will give our reasoning of selecting the exhaustive approach for the further analysis of our experimental data as well as the virtual dataset. Also the development of the classification rules has been discussed here in this chapter. I have also explored how these rules can benefit further development of new materials as well as how they can affect future applications in the field of wear resistance applications.

5.1 Alternative Feature Selection

This is used to select features relevant to a particular application. It helps in removing irrelevant and/or redundant data, hence improves the data quality and makes mining algorithms work faster on larger sized data. It enhances the comprehensibility of mined results as well. The feature selection ensures that the data fed to the data mining algorithm applications, is performed effectively and efficiently.

We used principal component analysis for feature selection in our previous chapters but for the analysis in this chapter we will use them but will include another feature selection method known as CFS subset evaluation method [1,2].

It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Also exhaustive

search was done for this evaluation, as it performs an exhaustive search through the space of attribute subsets starting from the empty set of attributes and reports the best subset found.

5.2 Results of Feature Selection

The results for hardness were similar to the results of the PCA analysis and following were the important materials:

- Larger Covalent Radius (anion+cation)
- Larger first Ionization Potential (anion+cation)
- Larger specific heat (anion+cation)
- Larger atomic radius (anion+cation)
- Larger atomic radius of cation.
- Larger density of an anion.
- Larger Molar Volume of anion.
- Larger pseudo potential core radii (anion+cation)

For friction the important attributes through CFS subset evaluation were not quite similar to that of PCA, following were the results

- Larger Valence electron anion
- Larger Boiling point (cation+anion)

- Larger Molar Volume cation
- Larger Heat of vaporization cation

For hardness the results were derived using these descriptors and came up with the apriori algorithm rules using classifiers while for friction the results of CFS subset evaluation gave no results even on decreasing the support and confidence to the minimum possible values. So to derive friction coefficient classification rules the important descriptors of the PCA analysis were considered. Hence conclusion of rules using these important descriptors was done for this analysis.

5.3 Heuristic and Exhaustive Search

There are two kinds of approaches in a data mining algorithms for searching rules in a data set, heuristic and exhaustive approach. Considering the need of finding *all* possible rules in the dataset to get useful insights into the correlation of important attributes to the properties of interest we decided to apply classification using association rules in our dataset. This is the kind of algorithm on which extensive research has been done in the data base community on learning rules using exhaustive search under the name association rule mining, as many existing classification and rule learning algorithm in machine learning mainly use the heuristic or greedy search to find a subset of regularities.

The issue with the heuristic approach is that they aim to find only a subset of regularities that exists in the data to form a classifier. In heuristic approach the covered examples are

either deleted or their weights are reduced for further formation of rules, this may hence not reflect the true regularities in the data and many high quality rules may not be found. Through apriori algorithm we use exhaustive search to find all rules in data that satisfy the user specified minimum support (5%) and minimum confidence (has been varied).

The aim here was to get all the rules as mentioned above and so the expectation was that the results will give way too many number of rules from which we might extract the important ones. As the rules with greater confidence rules will be very obvious and rules with less confidence can also have some value to see the unpredictable or new side of it. Also this kind of rules formation has not been done before in the material science field so the rules with higher confidence can also give useful insights. But surprisingly we got only six rules with a reasonable confidence percentage; these have been discussed later in the chapter.

5.4 Apriori Algorithm and the Methodology of class association rules

It is used for the classification of the reduced data set. Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules and that is what I have focused on in my research work.

The major strength of this system is that they are able to use the most accurate rules for classification because their rule learners aim to find all rules. This explains their good performance in general. However they also have weaknesses like they use only a single

minimum support value in rule generation, which can be inadequate sometimes for unbalanced class distributions. It generates all rules in two steps

1. Find all the frequent item sets that satisfy minimum support (*5%*)

2. Generate all the association rules that satisfy minimum confidence using the frequent item sets. (Varied from 100% to 70%)

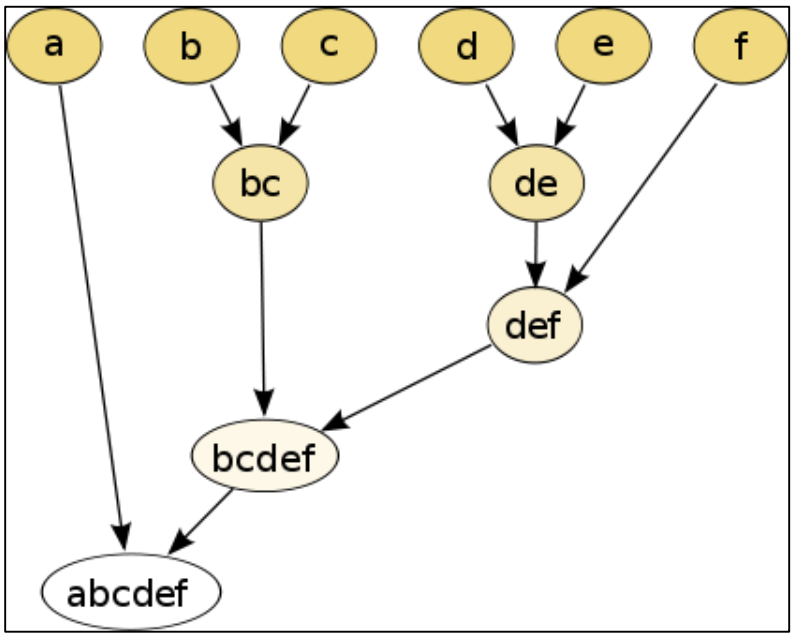

Fig 5.1 This figure explains the different levels of item sets and how the mining is performed. Here k=5.

To explain the concept further let's see what an itemset is? An itemset is a set of items. A frequent itemset is an itemset that has support above minimum support that is defined by the user. Mining of frequent itemsets is done in a level-wise fashion. Let k-itemset denote an itemset of k items. At level 1, all frequent 1-itemsets are found. At level 2, all frequent

2-itemsets are found and so on.
If an itemset is not frequent at level k-1, it is discarded, as any addition of items to the set cannot be frequent (this is called downward closure property).
At level k, all potentially frequent itemsets (candidate itemsets) are generated from frequent itemsets at level k-1. To determine which of the candidate itemsets are actually frequent, the algorithm goes through the data to count their supports.
After all frequent itemsets are found, it generates rules, which is relatively simple.

Mining association rules for classification from a continuous data set is done by taking a classification data set is in the form of relational table, which is described by a set of distinct attributes (discrete and continuous). A point that should be noted here is that association algorithm cannot be performed on a continuous dataset. So we first discretize each continuous attributes. After discretization, we can then transform each data record to a set of (attribute, value) pair of an item [3]

To generate all rules for classification we also need to make some modifications to the Apriori algorithm because a dataset for classification has a fixed target, the class attribute. Thus we only need to generate those rules X belonging to Ci, where Ci is a possible class. We call such association rules Class Association rules (CARs).

$X \longrightarrow c_i$ where c_i is a possible class. Three classes were defined for the data for each engineering property, friction as well as hardness. Class 1 was defined as hardness of a material better than the other, class 2 was defined as equal or not and class 3 was defined as hardness of compound A better than twice of hardness of compound B. Similarly three

classes for friction data set were also defined. After running the algorithm on the data set only class 1 attributes gave good results. There were three results found for hardness and six for friction coefficient, three best rules were selected for friction. The total six rules have been presented below with their proper explanation. Also the terms confidence, support and lift are to understood before exploring the rules. By confidence we mean conditional accuracy, hence more confidence the better is the accuracy of the rule, at the same time rules with less accuracy can also be useful insights as they can reveal the territory which has never been explored before. By support we mean the count of number of results which support the rule, the minimum the support the better as we don't want our rules to be biased and lift is the measure of correlation, if lift is greater than 1 it has positive correlation and if it is below than 1 it has negative correlation. We have targeted results that have positive correlation so the more the lift is the better the results is.

Rule 1

Rule	1
A>B (Specific Heat)	yes
B>A (Molar Volume anion)	yes
B>A (Atomic Radius)	yes
B>A (density of anion)	yes
B>A (pseudo potential core radii)	yes
Hardness	True

Rule 1. (Continued)

Confidence	84%
Support	0.2432
Lift	3.4539

This rule states that if we are comparing two compounds let's say A and B and if A has a better specific heat than compound B and B has better molar volume of anionic part and better atomic radius, density of anion and pseudo potential core radii than the hardness of A will be better than B 84% of the time with a support of .2432 and lift as 3.4539

Rule 2

Rule	2
A>B (First Ionization Potential)	yes
B>A (Molar Volume anion)	yes
B>A (Atomic Radius of cation)	yes
B>A (Covalent radius)	yes
B>A (pseudo potential core radii cation)	yes
Hardness (A > B)	True
Confidence	92%

Rule 2 (Continued)

Support	0.2432
Lift	3.7797

This rule states that if we are comparing two compounds let's say A and B and if A has a better First Ionization Potential than compound B and B has better molar volume of anionic part and better atomic radius of cation, covalent radius and pseudo potential core radii sum of cation than the hardness of A will be better than B 92% of the time with a support of .2432 and lift as 3.7797.

These two above rules have very high confidence and lift values and at the same time reasonably low support.

Rule 3

Rule	3
A>B (First Ionization Potential)	yes
A>B (Specific heat)	yes
B>A (Atomic Radius of cation)	yes
B>A (pseudo potential core radii)	yes
Hardness (A > B)	True

Rule 3 (Continued)

Confidence	81%
Support	0.4324
Lift	1.873

This rule states that if we are comparing two compounds let's say A and B and if A has a better First Ionization Potential and specific heat than compound B and B has better atomic radius, and pseudo potential core radii sum than the hardness of A will be better than B 81% of the time with a support of .4324 and lift as 1.873.

Further let's explore the top three rules of friction. For friction rules the desired result is minimum friction hence the following rules also indicate the combinations in the similar direction.

Rule 4

Rule	4
A<B (Melting Point anion)	yes
A<B (Boiling Point anion)	yes
A<B (pseudo potential core radii anion)	yes
Friction (A<B)	yes

Rule 4 (Continued)

Confidence	71%
Support	0.575
Lift	1.2347

This rule states that if we are comparing two compounds let's say A and B and if A's anionic part of the compound has smaller melting point, smaller boiling point and smaller pseudo potential core radii sum than compound B's anion than the friction of A will be less than B 71% of the time with a support of .575 and lift as 1.2347.

Rule 5

Rule	5
A<B (Density of anion)	yes
A<B (pseudo potential core radii anion)	yes
Friction (A<B)	yes
Confidence	71%
Support	0.57
Lift	1.2456

This rule states that if we are comparing two compounds let's say A and B and if A's anionic part of the compound has smaller density and smaller pseudo potential core radii sum than compound B's anion than the friction of A will be less than B 71% of the time with a support of .57 and lift as 1.2456.

Rule 6

Rule	6
A<B (Boiling point)	yes
A<B (pseudo potential core radii anion)	yes
Friction (A<B)	True
Confidence	71%
Support	0.59
Lift	1.19730

This rule states that if we are comparing two compounds let's say A and B and if A's anionic part of the compound has smaller boiling point and smaller pseudo potential core radii sum than compound B's anion than the friction of A will be less than B 71% of the time with a support of .59 and lift as 1.19730.

To demonstrate all the rules in a graphic form and showing the importance of each is what has been demonstrated in the following figure.

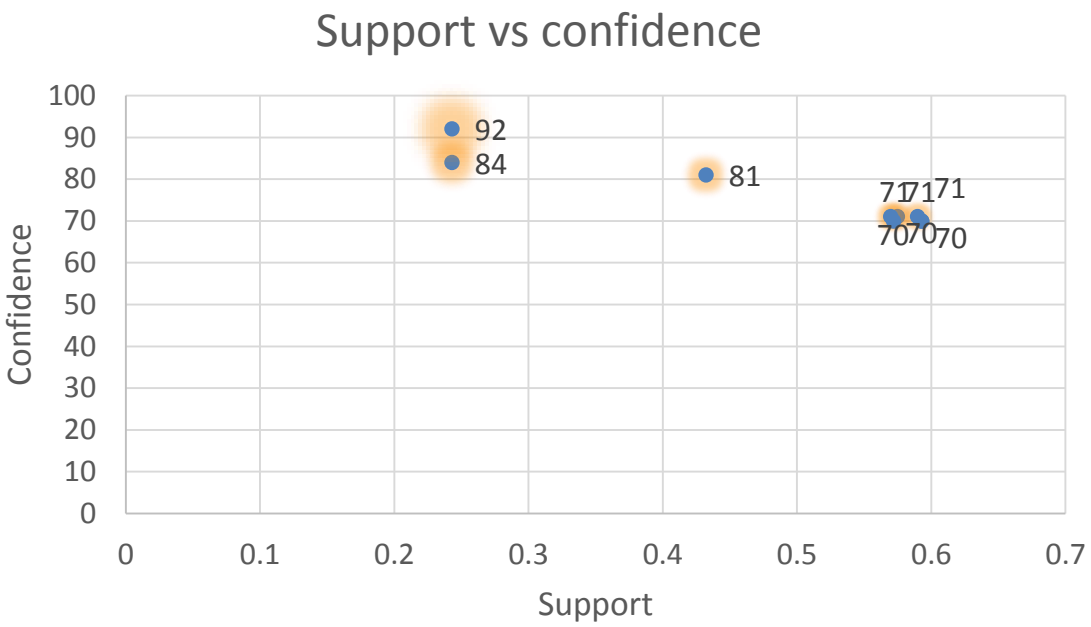

Fig 5.2 All nine rules demonstrated in a graphic form placed on a confidence vs. support two dimension plane. The more highlighted points are the better is the value of lift and hence the better is the rule.

If the numeric predictions are over fitting then to get some information out of such data where the independent variables are higher than the data points, we can perform classification algorithms as they provide the little nuggets of insights help in having a complete predictive model.

5.5 Depicting the above results in a decision making process

After coming up with the classification rules and applying it to the virtual database, a dataset was derived showing each compound's comparison with the other and describing which one is better and why? So for each comparison the dataset describes which rules make one compound better than the other. This data can then be arranged in a flow chart form with various decision-making algorithms and has been explored below.

Taking nine compounds from our dataset and after constructing an excel sheet by comparing all 9 compounds with each other and describing for each set how one is better than other has been described in the following table using the classification rules. The chart above the diagonal line shows how the column compounds are better than the compounds in the row and the chart below the diagonal line shows how the compounds in the row are better than the compounds in the column. For example we have explored how AlF3 is better than Al2O3 and how Al2O3 is better than AlF3 and have then made a decision.

Table 3: Comparing each of these nine compounds with another to see what rules of classification makes one better than the other.

Compound	AlF3	Al2Se3	AlAs	Al2Te3	Al2S3	AlCl3	AlBr3	SiF4	SiSe2
AlF3	-	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	3	3	none	**1,2,3,4,5,6**
Al2Se3	none	-	4,5,6	none	None	none	none	none	**none**
AlAs	none	none	-	3	None	none	none	none	**none**
Al2Te3	none	none	4,5,6	-	None	none	none	none	**none**
Al2S3	none	3	1,2,3,4,5,6	3	-	none	none	none	**none**
AlCl3	none	1,2,3,4,5,6	4,5,6	1,2,3,4,5,6	4,5,6	-	3	none	**1,2,3,4,5,6**
AlBr3	none	1,2,3,4,5,6	4,5,6	1,2,3,4,5,6	4,5,6	none	-	none	**1,2,3,4,5,6**
SiF4	5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,5,6	1,2,3,5,6	-	**1,2,3,4,5,6**
SiSe2	**none**	**5,6**	**5,6**	**5,6**	**none**	**none**	**none**	**none**	**-**

By simply depicting the above results in the flow chart form we get the following figure. In this chart we can see that SiF4 is better than all the other compounds and how it is better

than AlF3 (the next best compound with minimum friction and maximum hardness) has been specifically mentioned. For example SiF4 is better than AlF3 in terms of friction rule 5 and 6, in terms of hardness they cannot be differentiated. So by using this simple technique one can evaluate all 135 compounds after comparing each compound and formulating an excel sheet and then running an algorithm for the above mentioned analysis. Also we can see in the chart that AlAs and Al2Te3 are contradicting each other one is better than other in some respect, in such cases a further understanding of decision criteria and preference setting has to be done by the decision maker. But in the following case we are getting a straight forward answer of a best compound out of the lot of 9 compounds which is SiF4.

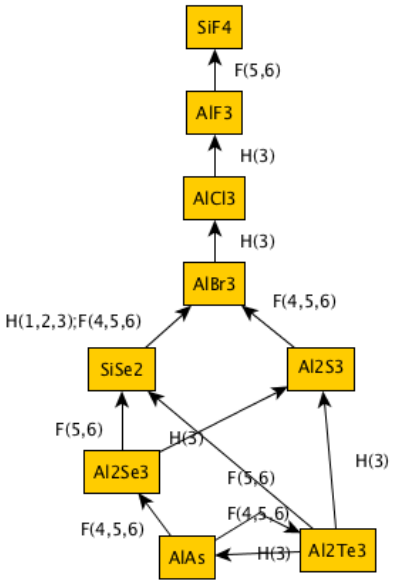

Fig 5.3 The flow chart depicting the results of comparison based on the classification rules.

Also the comparison of the 9 compounds highlighted in chapter 4 which have been selected as prospective compounds for wear resistance application were diagnosed by the rules we developed through apriori.

Table 4: Comparing each of these nine highlighted compounds with another to see what rules of classification makes one better than the other.

Compound	SiF4	SnF4	ZrF4	TiF4	MoF4	VF5	BaO	PbO	AgO
SiF4	-	1,3,5	1,3,4,5,6	3,4,5,6	3,4,5,6	4,5,6	1,2,3,4,5,6	1,2,3,5,6	1,2,3,4,5,6
SnF4	none	-	4,5,6	4,6	4,5,6	4,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6
ZrF4	none	none	-	none	none	none	1,2,3,4,5,6	1,2,3,5,6	1,2,3,4,5,6
TiF4	none	none	1,4,5,6	-	4,5,6	none	1,2,3,4,5,6	1,2,3,5,6	1,2,3,4,5,6
MoF4	none	none	none	none	-	none	2,3	1,2,3,5,6	1,2,3,4,5,6
VF5	none	none	1,3,4,5,6	3,4,6	1,3,4,5,6	-	1,2,3,4,5,6	1,2,3,5,6	1,2,3,4,5,6
BaO	none	none	none	none	none	none	-	none	none
PbO	none	none	none	none	none	none	none	-	4,6
AgO	none	none	none	none	none	none	none	none	-

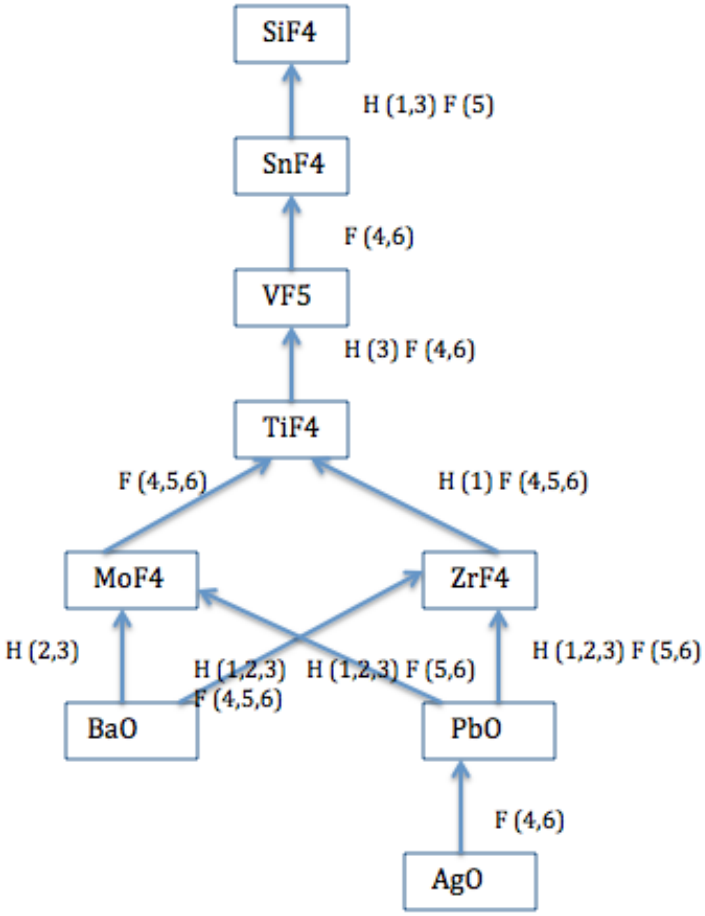

Fig 5.4 The flow chart depicting the results of comparison based on the classification rules.

SiF4 turned out to be best amongst the 9 prospective compounds and almost all the results were in consistency with each other.

As can be seen in the next figure, with the help of classification rules we can easily arrange the materials in their order of importance as it would not have been possible by the previous method because to decide tradeoffs in this case was difficult. Again as described in the previous example to decide tradeoffs between MoF4 and ZrF4 is not possible through rules as they are incomparable. BaO, PbO and AgO lie at the bottom as when friction and

hardness combination is looked on they are the lowest. These results indicate that through both the methods used for small data sizes, we can check whether the prediction modelling is overfitting or not as it is based on the transformed space of the dataset whereas the other method uses original feature space. Given the results are same, hence there is a high possibility of no overfitting in the prediction model.

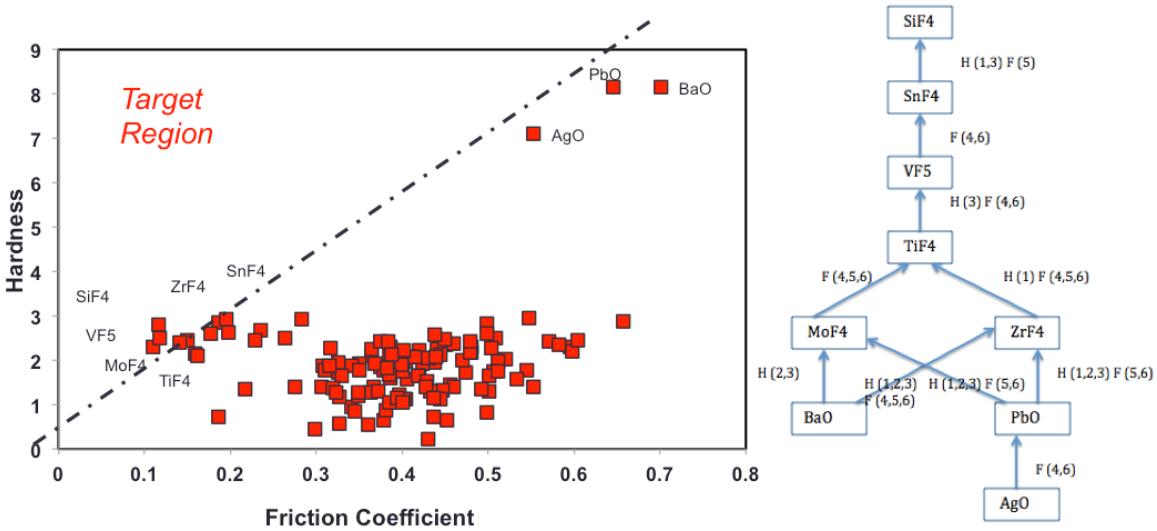

Fig 5.5 The result comparison of both the methods. Figure on the left is the result of a prediction model whereas the result on the right is a result of the classification rules.

There are several properties that come into factor when we analyze wear resistance of a material like friction coefficient, hardness etc. For a given set of materials all these properties are analyzed and the selection of a best material amongst those is done. Selection of the best material involves ranking of materials. There are different Multi Criteria Decision Analysis techniques that are used for the ranking of the materials in the given

data set. The above description shows how it can be achieved in one way. Hence the next student who might join the research group can further explore this. This aspect has been considered for future work and is not fully explored in this part of the thesis. But realizing how the rules can help in making a well-informed decision in material selection using this technique was the intension behind explaining the methodology.

5.5 References

1. H. Liu and H. Motoda. Feature selection for knowledge discovery and data mining, Kluwer Academic, Massachusetts, 2000.

2. John, G., Kohavi, R., and Pfleger, K. (1994). Irrelevant feature and the subset selection problem. In Machine learning: Proceedings of the Eleventh International Conference, pages 121-129. Morgan Kaufmann Publisher.

3. Bing Liu, Yiming Ma, and Ching-Kian Wong. Classification using association rules: weaknesses and enhancement.

CHAPTER 6

CONCLUSIONS

This thesis developed a new hybrid informatics approach for identifying target materials for further exploration. This approach connected dimensionality reduction, attribute selection, prediction and association mining approaches, utilizing and linking aspects of each for a unified design strategy. We applied this approach to ceramic wear resistant materials for improving hardness and friction coefficient to expand their applicability to high temperature environments. The particular challenges addressed through this approach include small existing knowledge base and high data uncertainty.

In this work, the knowledge base of binary ceramics for wear applications has been increase by five times what was previously known. This increase is particularly significant given the difficulty associated with obtaining the wear data, which has resulted in the small data knowledge base. Further, the work emphasized modeling robustness over accuracy whenever the trade-off was needed. The reason for this was to ensure data was not being over-fit. This challenge arises due to the small data that was input. By using the comparison of two different approaches, one data driven and the other working on the transformed dataset and by getting similar results we have ensured that any materials that we identify as having promising characteristics are highly likely to have those characteristics. By avoiding over-fitting or sensitivity to outliers and selecting materials on extracted physics, we have enhanced the robustness of our material selections. The usage of decision theory was applied and described in this thesis, and this represents a

promising area to be further coupled with the hybrid methodology developed and utilized here.

Suggestions for future work are to increase the system complexity by introducing multi-objective functions, such as surface tension, melting point etc., as they also play an important role in affecting the wear resistance of a material. Also developing efficient qualitative multi-attribute decision theory algorithms to find optimal choice amongst the expanded data. These methodologies are totally new in the field of materials science. The models, data and experiments, some of which are discrete and some are based on differential equations; can use these techniques to interpret the model.